Weekly Report 11/03/2013

The data mining and cleaning project --- more thoughts

Basic diagnostics in data debugging perspective

- single field --- missing value, format mismatch, value distribution, duplication
- multiple fields
    - entity vs. entity --- correlation, duplicate relation
    - entity vs. time --- temporal distribution/trend
    - entity vs. geolocation --- spatial distribution
    - field vs. subfield --- hierarchical feature?
- multiple schemas --- schema mismatch (validation of one-to-one mapping)

Verification shares the same (visualization) methods with diagnostics expect in a data display perspective instead of data debugging perspective.

Basic transformation functions of filter unit

- renaming --- which renames the name of the data field.
- reformat --- which changes the format of the data field.
- adding --- which adds data fields to the schema.
- deleting --- which deletes data fields from the schema.
- splitting --- which splits one data fields to two or more fields.
- merging --- which merges two data fields in one schema or across schema into one field
- extracting --- field extracting derives values (substrings for example) from one data field and creates a new field with it; schema extracting derives multiple data fields into new schema.
- schema mapping --- which maps data fields in one schema to another schema, or to other fields in the same schema. It may result in new data fields created in the original schema, or new schema created to represent the mapping, or original schema replaced by the new schema.

System Implementation (still open)

system architecture (B/S)

- web rendering: D3.js maybe
- application server: Python (Django framework) with C++ computation modules integrated
- HTTP server: NginX (event-driven, high-performance, fast response) or Apache (easy to use, lots of documents and resources)

data construction: not yet

database: mysql

schema language: not yet

visual display: D3.js

I'm trying to set up the Django web architecture with C++ module integration.

I've sent a proposal draft to Dr. Ebert, but didn't get any response yet.

The VASA project

Get the project code and dataset from two project members. Since four universities are collaborating together on this project, the data protocol is still messed up. I'm getting to know the dataset by doing some basic data cleaning and parsing job.
Since it's a simulation project, we're also doing surveys on simulation visualization. I've read the paper *World Lines*. The timeline with branches design is a good reference.

The sound project

This project hasn't officially started yet. There are just some basic ideas and background. The sound processing part would mainly be done by another team. Currently we have to set up a portal website for users to upload and explore the sound database.

Not much work has been done in this week. Because I have just talked to Dr. Ebert this Monday and get to know the projects I can take part in. Also, due to the VASA project program problem, I installed operating system several times in the lab… That was really frustrating.

Future Work

1. Django web architecture study. Startups on web UI.
2. More implementation decomposition of the data cleaning project. Start core function implementation as soon as possible.
3. Follow the VASA project schedule. More papers.